

# Flexible, symmetry-directed approach to assembling protein cages

Aaron Sciore<sup>a</sup>, Min Su<sup>b</sup>, Philipp Koldewey<sup>c</sup>, Joseph D. Eschweiler<sup>a</sup>, Kelsey A. Diffley<sup>a</sup>, Brian M. Linhares<sup>b</sup>, Brandon T. Ruotolo<sup>a</sup>, James C. A. Bardwell<sup>c,d,e</sup>, Georgios Skiniotis<sup>b,d</sup>, and E. Neil G. Marsh<sup>a,d,1</sup>

<sup>a</sup>Department of Chemistry, University of Michigan, Ann Arbor, MI 48109; <sup>b</sup>Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109; <sup>c</sup>Department of Molecular, Cellular and Developmental Biology, University of Michigan, Ann Arbor, MI 48109; <sup>d</sup>Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109; and <sup>e</sup>Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI 48109

Edited by David Baker, University of Washington, Seattle, WA, and approved June 10, 2016 (received for review April 15, 2016)

**The assembly of individual protein subunits into large-scale symmetrical structures is widespread in nature and confers new biological properties. Engineered protein assemblies have potential applications in nanotechnology and medicine; however, a major challenge in engineering assemblies de novo has been to design interactions between the protein subunits so that they specifically assemble into the desired structure. Here we demonstrate a simple, generalizable approach to assemble proteins into cage-like structures that uses short de novo designed coiled-coil domains to mediate assembly. We assembled eight copies of a C<sub>3</sub>-symmetric trimeric esterase into a well-defined octahedral protein cage by appending a C<sub>4</sub>-symmetric coiled-coil domain to the protein through a short, flexible linker sequence, with the approximate length of the linker sequence determined by computational modeling. The structure of the cage was verified using a combination of analytical ultracentrifugation, native electrospray mass spectrometry, and negative stain and cryoelectron microscopy. For the protein cage to assemble correctly, it was necessary to optimize the length of the linker sequence. This observation suggests that flexibility between the two protein domains is important to allow the protein subunits sufficient freedom to assemble into the geometry specified by the combination of C<sub>4</sub> and C<sub>3</sub> symmetry elements. Because this approach is inherently modular and places minimal requirements on the structural features of the protein building blocks, it could be extended to assemble a wide variety of proteins into structures with different symmetries.**

coiled coils | protein design | native mass spectrometry | analytical ultracentrifugation | cryoelectron microscopy

**T**he assembly of individual protein subunits into large-scale structures, often from only one or a few types of protein monomer, is widespread in nature; examples include viral capsids, multienzyme complexes, and intracellular storage compartments (1–4). These protein assemblies are generally characterized by a high degree of symmetry. An important consequence of the assembly process is the emergence of more complex biological properties; well-studied examples include the dynamic polymerization of actin (5) and tubulin fibrils (6) and the GroEL/GroES, which is a protein chaperone complex (7). In their assembled form, the basic ATPase activity inherent to each of these proteins is harnessed toward the more complex tasks of motility and protein refolding, respectively. Consequently, there is significant interest in the fields of synthetic biology and nanotechnology in designing novel self-assembling proteins and adapting natural protein assemblies for a range of applications broadly encompassing nanomedicine and materials science (4, 8–13).

Early work by Yeates and coworkers (14, 15) recognized that the principles of symmetry, often used in the design of inorganic materials, could be exploited to design either discrete, cage-like protein assemblies or extensive networks in one, two, and three dimensions. An important realization was that a large number of complex symmetries could be generated from only two distinct symmetry elements (for a protein, these must be rotational symmetries specified by its quaternary structure), provided the

orientation of the symmetry axes with respect to each other could be carefully controlled. These principles have now been quite widely applied to design both protein cages and protein networks (16–22). The principal challenge to researchers has been to design new interactions between the protein subunits that promote assembly in the desired geometry, and, in particular, to align the angle between symmetry axes correctly. A variety of strategies have been used to facilitate assembly; these include genetically linking two protein interaction domains (14, 23, 24), the use of bifunctional ligands and metal ions to coordinate proteins (19, 21, 22, 25), and the computational design of new protein–protein interfaces (16, 17).

Despite significant progress, the design of protein systems that assemble into well-defined architectures remains a challenging goal. Whereas genetically linking two protein interaction domains together is easy to accomplish, it has proven hard to achieve the necessary degree of control over the orientation of the proteins. In only a few cases has this approach yielded assemblies that are sufficiently homogenous to characterize crystallographically (26, 27). More often genetically linking protein interaction domains result in polydisperse protein assemblies (21, 22, 28–31); these are hard to characterize and are limited in their potential utility.

More recently, the computation redesign of protein–protein interfaces has met with some impressive successes, leading to the construction of rigid protein cages that could be characterized crystallographically (16, 17). However, this protein redesign is computationally intensive and requires very precise control of the protein–protein interfaces to successfully direct assembly. The precision needed to successfully redesign protein–protein interfaces limits the number of proteins amenable to this approach and requires that many designed variants be experimentally screened to

## Significance

**The ability to organize biological molecules into new hierarchical forms represents an important goal in synthetic biology. However, designing new quaternary interactions between protein subunits has proved technically challenging and has generally required extensive redesign of protein–protein interfaces. Here, we demonstrate a conceptually simple way to assemble a protein into a well-defined geometric structure that uses coiled-coil sequences as “off-the-shelf” components. This approach is inherently modular and adaptable to a wide range of proteins and symmetries, opening up avenues for the construction of biological structures with diverse geometries and wide-ranging functionalities.**

Author contributions: A.S., M.S., P.K., B.T.R., and E.N.G.M. designed research; A.S., M.S., P.K., J.D.E., K.A.D., and B.M.L. performed research; A.S., M.S., P.K., B.T.R., J.C.A.B., G.S., and E.N.G.M. analyzed data; and A.S., M.S., P.K., B.T.R., J.C.A.B., G.S., and E.N.G.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

<sup>1</sup>To whom correspondence should be addressed. Email: nmarsh@umich.edu.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606013113/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606013113/-DCSupplemental).

identify well-folded assemblies. Moreover, the extensive reengineering of the protein surface that is often needed to construct the interface may negatively impact the biological activity of the designed protein.

We aimed to develop a general approach to designing protein assemblies that is largely independent of the structural details of the engineered protein and that does not require the orientation of the symmetry axes to be explicitly specified. Here, we describe a strategy for assembling a trimeric protein into an octahedral cage using a small de novo designed, parallel four-helix bundle coiled-coil domain that is genetically fused to the C terminus of the protein through a short, flexible linker. The structure of assembly is primarily specified by the symmetry of the coiled-coil domain. We show that, despite the flexibility of the linker, the resulting protein cage adopts a well-defined structure and is highly homogeneous.

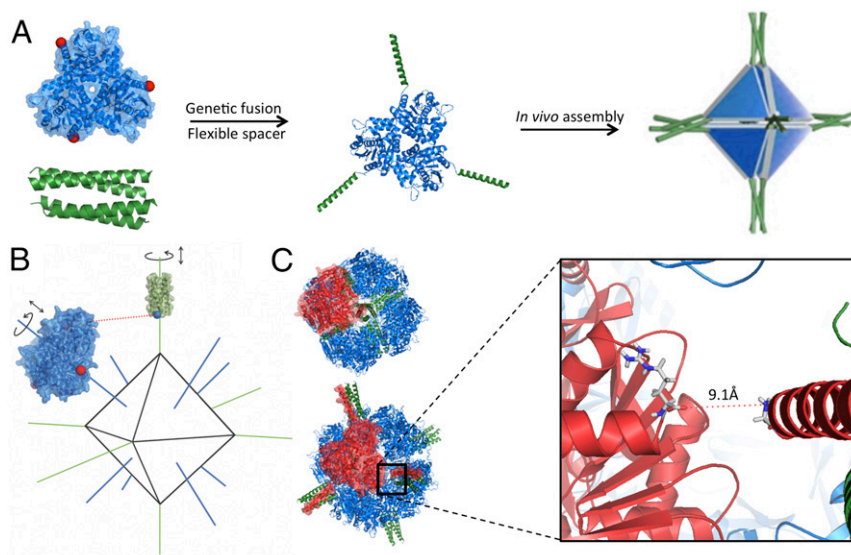
## Results

**Design Approach.** In our design approach, we sought to develop a flexible, modular strategy in which the protein building block and the coiled-coil domain function independently but, when genetically linked together, assemble into a single structure of the desired symmetry. In general, attempts to design protein assemblies have focused almost exclusively on combining trimeric ( $C_3$ -symmetric) proteins with dimeric ( $C_2$ -symmetric) proteins, as these are common quaternary structures (10). The combination of  $C_3$  and  $C_2$  symmetry elements occurs in multiple point groups, so many geometries are compatible with assemblies made up of these symmetry elements. In contrast, the combination of  $C_3$  and  $C_4$  symmetry elements is unique to the octahedral point group. Therefore, we attempted to construct an octahedral protein cage based solely on combining proteins with these rotation symmetries and without explicit orientation constraints.

We surveyed several trimeric proteins in the Protein Data Bank (PDB) and selected, as a test case, a trimeric esterase, PDB ID 1ZOI (32). In this esterase, the C terminus is oriented toward the apex of the triangle formed by the  $C_3$ -symmetric protein, positioning it in approximately the right place to facilitate addition of

the  $C_4$ -symmetric domain. (Fig. 1A). Natural,  $C_4$ -symmetric proteins are rare, as most tetrameric proteins adopt a pseudo- $D_2$  “dimer-of-dimers” symmetry. Therefore, we used a de novo designed coiled-coil protein as the  $C_4$  component. Coiled coils are among the simplest and best-understood protein–protein interactions (33). As such, there are a large number of well-characterized designs available as “off-the-shelf” components for use in protein engineering applications, including dimeric, trimeric, tetrameric, pentameric, and hexameric designs in both parallel and antiparallel forms (34–36). A further advantage is that the strength of the coiled-coil interaction can easily be manipulated by varying the number of heptad repeats. For our purposes, we selected a parallel, four-helix coiled coil in which the tetrameric arrangement is specified by four repeating heptads in which Leu and Ile are present at the “a” and “d” positions of the canonical heptad (37); the crystal structure of this protein, PDB ID 3R4A (37), shows that it possesses close to perfect  $C_4$  symmetry.

To determine the approximate minimum length of flexible linker needed to connect the C terminus of the  $C_3$  protein with the N terminus of the  $C_4$  coiled coil, we aligned the  $C_3$  axis of the esterase and the  $C_4$  axis of the coiled coil along the  $C_3$  and  $C_4$  axes, respectively, of the octahedral point group. Using a search algorithm implemented in the program Rosetta (38), the angle of rotation of each protein about its symmetry axis and its distance from the origin were allowed to vary in a symmetrically constrained manner. The distance between the two termini was minimized, discarding any configurations with steric clashes (defined as any intersubunit backbone atom distances shorter than 4 Å) (Fig. 1B). The modeling indicated that the coiled coils could either point inward or outward. (The inward-pointing orientations were examined by negatively translating the coiled-coil coordinates along the symmetry axes indicated in Fig. 1B. This orientation is feasible because the vertices of the trimeric esterase don’t pack together perfectly, leaving sufficient space for the coiled-coil domain to point inward while still maintaining a compact structure.) Either orientation yielded a similar minimum distance between the termini of the esterase and coiled coil of  $\sim 9.1$  Å that could, in principle, be



**Fig. 1.** Design of a self-assembling octahedral protein cage. (A) Structures of the trimeric esterase (PDB 1ZOI) (C termini of the esterase are indicated by red spheres) and the tetrameric coiled coil (PDB 3R4A) used in the design. (B) Minimization of linker distance compatible with octahedral geometry. The proteins were arrayed along the  $C_3$  (blue line) and  $C_4$  (green line) symmetry axes, and the distance between the N terminus of the coiled coil and the C terminus of the esterase (dashed red line) was minimized by symmetrically varying the rotation of the proteins about the symmetry axes and their radial distance while avoiding steric clashes. (C) Distance-minimized structures were found to be compatible with the coiled-coil domains either facing inward (top structure) or outward (bottom structure) with a minimum interterminus distance of  $\sim 9.1$  Å.

bridged by a minimum of three amino acid residues (Fig. 1C). PDB files of the models are provided as [Datasets S1](#) and [S2](#).

Based on the modeling, we constructed three synthetic genes ([Table S1](#)) in which the C terminus of the trimeric esterase was genetically fused to the N terminus of the tetrameric coiled-coil domain through a flexible linker sequence comprising two, three, or four glycine residues that potentially could span between 6 Å and 12 Å. We refer to these designs as Oct-2, Oct-3, and Oct-4, respectively.

**Initial Characterization of Protein Cage Designs.** The genes encoding Oct-2, Oct-3, and Oct-4 were overexpressed in *Escherichia coli*. Of the three designs, Oct-2 and Oct-4 expressed as soluble proteins, whereas, for reasons that are unclear, Oct-3 was produced only as inclusion bodies. (Oct-2 and Oct-4 were also observed to form inclusion bodies, but to a much lesser extent.) Oct-2 and Oct-4 were purified to homogeneity using an N-terminal His-tag by standard methods ([Fig. S1](#)) and were initially screened for their ability to assemble into discrete complexes using size exclusion chromatography (SEC) and native PAGE ([Fig. 2](#)). Oct-2 formed a heterogeneous mixture of assemblies that, by SEC, appeared to be too large to represent an octahedral cage, whereas Oct-4 appeared more homogeneous, assembling into a complex of approximately the correct size for an octahedral cage and judged to be nearly homogenous by native PAGE. We therefore selected Oct-4 for more detailed characterization by analytical ultracentrifugation (AUC), native electrospray ionization mass spectrometry (ESI MS), and negative stain and cryoelectron microscopy (cryo-EM).

**AUC of Oct-4.** Sedimentation velocity AUC provides a powerful method for analyzing macromolecules in solution and can provide detailed information on the number of species present and their hydrodynamic properties (39, 40). Oct-4 (0.2 mg/mL in 100 mM NaCl, 25 mM Hepes, 1 mM EDTA buffer, pH 7.5) was sedimented at  $94,350 \times g$ , and the sedimentation traces were analyzed by two-dimensional sedimentation spectrum analysis (2DSA) using the program Ultrascan (41); this is a model-independent analytical approach to fit sedimentation velocity traces to the Lamm equation that allows both the shape (frictional ratio) and molecular

mass distribution of macromolecular mixtures to be independently and reliably determined.

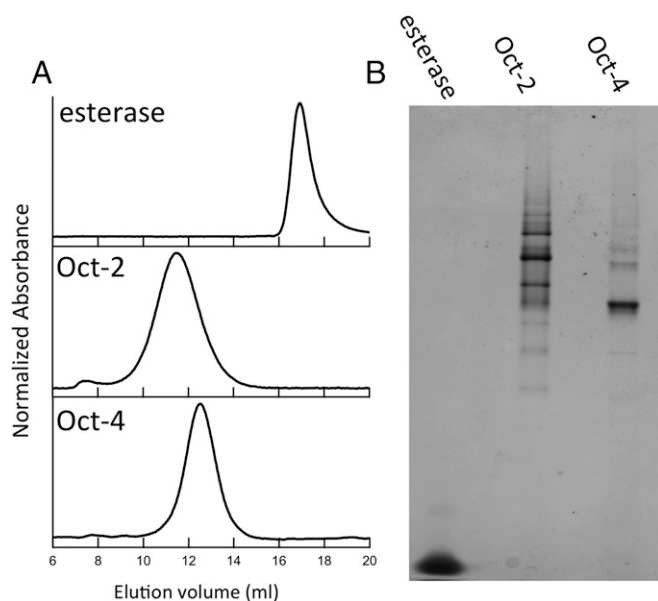
From this analysis, Oct-4 was found to comprise predominantly (~75%) a single hydrodynamic species ([Table S2](#)), in good agreement with native PAGE. The sedimentation coefficient ( $s_{20,w}$ ) and frictional ratio ( $f/f_0$ ) of this species were 17.5 S and 1.89, respectively ([Fig. 3A](#)). From these data, a molecular mass of  $886 \pm 14$  kDa was calculated, which is in good agreement with the expected mass of 854 kDa calculated for the assembly of 24 subunits into an octahedral cage. The frictional ratio is somewhat higher than expected for simple globular protein; this may be attributed to the porous nature of the cage, which would be expected to increase the interaction with the solvent. The  $f/f_0$  is within the range measured for other porous protein cages such as ferritin,  $f/f_0 = 1.3$  (4), and the E2 complex of pyruvate dehydrogenase,  $f/f_0 = 2.5$  (42).

We also undertook a 2DSA analysis of Oct-2 under the same experimental conditions. This analysis indicated that multiple species were present, with sedimentation coefficients ranging between 25 and 58 S and frictional ratios ranging between 1.0 and 1.1 ([Fig. S2](#) and [Table S3](#)), which is consistent with the formation of a range of compact globular assemblies.

**Native Mass Spectrometry of Oct-4.** Native ESI MS induces the desolvation and ionization of biological molecules under very mild conditions, allowing the masses of large, noncovalent protein assemblies to be determined (43, 44). Samples of Oct-4, ~1 mg/mL, were buffer-exchanged into 200 mM ammonium acetate buffer, pH 7.0, and analyzed by native ESI MS. Initial mass spectra were collected using gentle conditions, i.e., low in-source activation voltages, so as not to dissociate the complex. This method of collection produced a spectrum containing a single broad distribution of unresolved peaks centered around  $m/z$  12,000, characteristic of a very large, incompletely desolvated complex. By carefully increasing the in-source activation voltage, the complex could be desolvated sufficiently to resolve a population of discretely charge states ([Fig. 3B](#)), allowing the mass of the complex to be calculated. This sequence yielded a mass of  $887 \pm 5$  kDa for Oct-4, which is 5.5% larger than predicted for a 24-subunit assembly. The broad peak envelope and increased mass may be attributed to incomplete desolvation of the complex; this is commonly encountered in the analysis of large porous molecular complexes, which effectively trap solvent and buffer ions within their structures (43). We also observed signals centered on  $m/z$  11,200 ([Fig. 3B](#)) that correspond to a mass of  $757 \pm 7$  kDa. We assign this species to an Oct-4 form, having lost one esterase trimer from the intact complex, likely during the buffer exchange procedure.

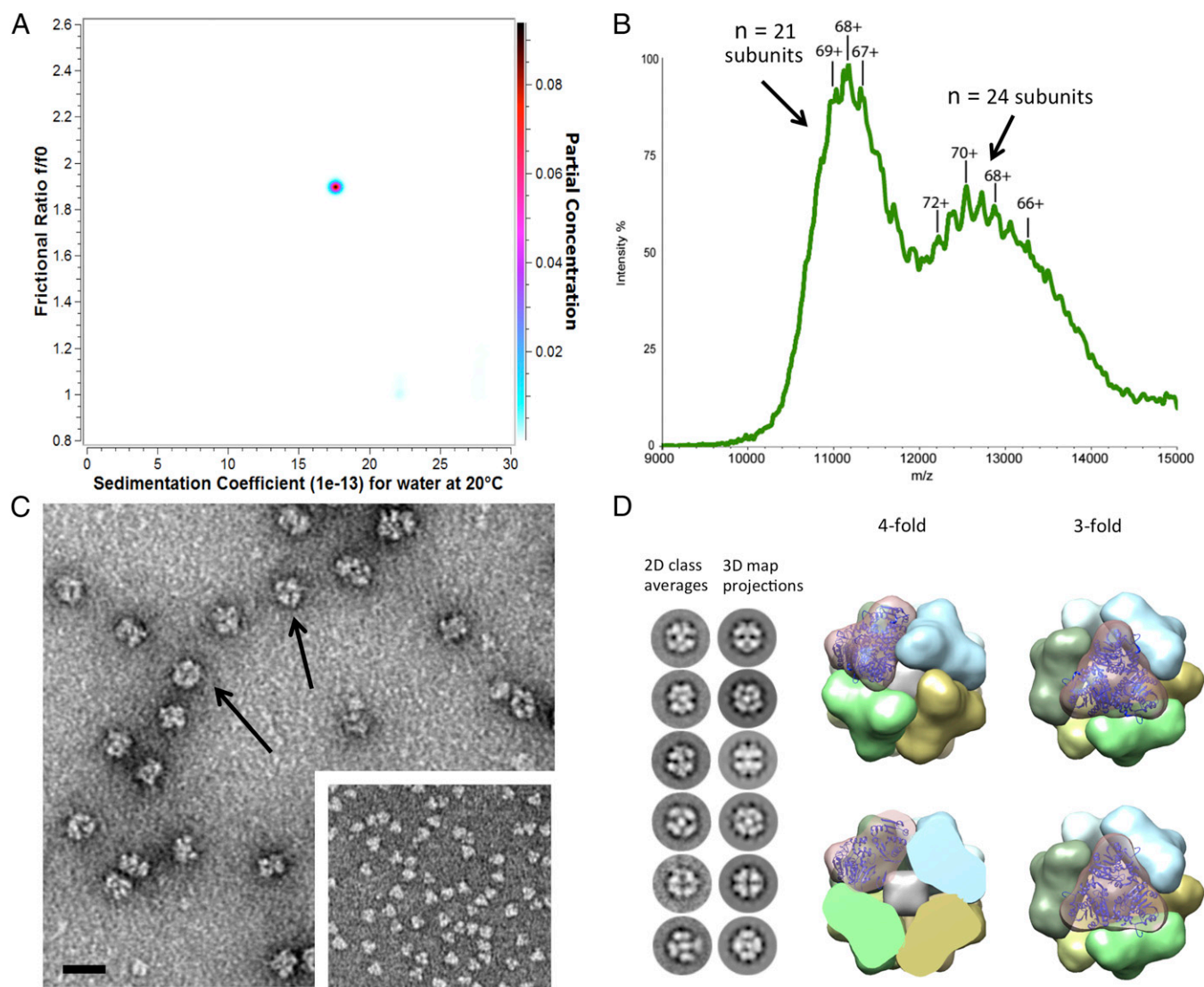
**Negative Stain and Cryo-EM of Oct-4.** Negative stain EM of Oct-4 ([Fig. 3C](#)) provided further evidence that this design adopts the intended octahedral architecture. EM images show that Oct-4 forms compact, globular structures of the diameter expected for an octahedral assembly (~18 nm), and, in some cases, the fourfold axis of symmetry is clearly discernable. In contrast, EM images of Oct-2 showed that, although it also forms compact globular structures, they are larger and more variable in size, and lack apparent symmetry ([Fig. S2](#)). We suspect that the heterogeneity present in Oct-2 is likely due to the linker sequence being too short to permit the components to assemble into the ideal octahedral geometry.

To further probe the architecture of Oct-4, we visualized preparations by cryo-EM and excised 44,856 particles for single-particle analysis. Particle images were subjected to reference-free classification and averaging using the program ISAC (45), thereby generating class averages ([Fig. 3D](#) and [Fig. S3](#)). Although the trimeric architecture of the esterase was clearly evident in many class averages, we did not observe any peripheral electron density that could be associated with the coiled-coil domains. The lack of electron density could be because these very small domains



**Fig. 2.** Initial characterization of Oct-2 and Oct-4. (A) SEC of Oct-4, Oct-2, and the unmodified esterase. (B) Native gel electrophoresis of Oct-4, Oct-2, and the unmodified esterase.





**Fig. 3.** Structural characterization of Oct-4. (A) Sedimentation velocity AUC of Oct-4. The protein sediments primarily (>75%) as a single, well-defined species with an appropriate weight and shape for a 24-subunit octahedron. (B) Native electrospray mass spectrum of intact Oct-4. The envelope of charge states centered at  $m/z$  12,600 corresponds to a species of  $M_r = 887 \pm 5$  kDa, whereas those centered at  $m/z = 11200$  corresponds to a species of  $M_r = 757 \pm 7$  kDa. The smaller species represents dissociation of one trimer from the octahedral complex under the conditions of the Native MS experiment. (C) Negative stain EM images of the particles formed by Oct-4. Arrows indicate particles where fourfold symmetry is apparent. (Scale bar, 20 nm.) (Inset) Negative stain EM of unmodified trimeric esterase. (D, Left) Representative 2D class-averaged images of Oct-4 and projections generated from the 3D electron density map. (Right) Reconstructed electron density for Oct-4 viewed along the fourfold and threefold axes with one esterase trimer shown modeled into the electron density. The lower images show a slice through the electron density.

are flexible and average out, but it may also suggest that the coiled coils face inward, toward the center of the cage. Consistent with this latter hypothesis, a number of the class averages show enhanced electron density at the center of the averaged particles that could reflect inward-facing coiled coils. Also, in many of the class averages, the protein cages appear distorted, which further suggests that the assembled complexes are conformationally flexible.

To better understand the cage structure, we used 34,980 particle projections belonging to the most well defined averages and calculated a low-resolution 3D cryo-EM reconstruction of Oct-4 with an indicated resolution of 17 Å. The symmetrically reconstructed map reveals the octahedral cage arrangement of distinct trimers representing the esterase, as confirmed by docking its crystal structure within the corresponding density (Fig. 3D). The low resolution of the EM map is consistent with the limited features presented in the class averages and again suggests that

the cages formed by Oct-4 are quite flexible, thereby leading to blurring of the averaged density. The reconstruction also contains a featureless region of additional electron density at the center of the cage. Although this additional electron density could be partly due to the octahedral symmetrization procedure used in the reconstruction, the volume of this central density suggests that it is part of the oligomeric assembly. The density could arise from the coiled-coil domains if they were oriented toward the interior of the cage.

**Catalytic Activity of Assembled Protein Cages.** An important consideration in the design and construction of protein cages is that the building block proteins should retain their biological activity when assembled. The esterase activity of the assembled protein cages Oct-2 and Oct-4 was compared with the unmodified trimeric esterase by following the hydrolysis of *p*-nitrophenyl acetate. The

specific activity of the unmodified esterase determined in 25 mM Hepes, pH 7.5, 100 mM NaCl, at 25 °C was  $54 \pm 4 \mu\text{M}\cdot\text{min}^{-1}\cdot\text{mg}^{-1}$ , whereas the specific activities of Oct-2 and Oct-4 were  $19.5 \pm 0.5$  and  $20 \pm 0.4 \mu\text{M}\cdot\text{min}^{-1}\cdot\text{mg}^{-1}$ , respectively.

The reason for the lower specific activities of the assembled proteins is currently unclear. It might be that assembly impedes substrate access to the active site, or that it imposes small distortions on the active site geometry or dynamics, both of which could lower activity. However, the retention of activity implies that the tertiary structure of the protein was not significantly altered by the assembly process.

## Discussion

Various studies have used symmetry-based methods for assembly of threefold symmetric proteins into octahedral and tetrahedral cages using other protein domains, bifunctional cross-linkers, metal ions, or designed protein interfaces to direct assembly (14, 16, 17, 19, 21–23, 25, 28–30). Common to these approaches has been the combination of  $C_3$  and  $C_2$  symmetry elements, which has required that the orientation of the two symmetry elements be carefully controlled to prevent the formation of heterogeneous assemblies. Here, we have shown that, by switching to a combination of  $C_3$  and  $C_4$  symmetry elements, it is possible to organize a protein into a geometrically well-defined, large-scale assembly without the need to explicitly specify the relative orientation of the two protein domains. To our knowledge, this is the first example of a designed protein cage that incorporates a  $C_4$ -symmetric element to mediate assembly.

It is worth noting that the flexible connection between the  $C_3$  and  $C_4$  symmetry elements, in principle, also permits larger structures of lower symmetry to be formed without violating the “ $4 \times 3$ ” valency rules. It is also possible that incompletely or incorrectly assembled structures could form that become kinetically trapped; this may explain the ensemble of larger assemblies that are formed by Oct-2, which possesses a shorter linker sequence. Indeed, some evidence for off-pathway assemblies was also evident in preparations of Oct-4, as evidenced by native PAGE (Fig. 2B), although SEC largely removed these during purification (Fig. S1).

We envisage that the coiled-coil domains act like “twist ties” to hold the esterase trimers in a flexible octahedral configuration. As such, the assembly process is, in principle, independent of the structural details of the protein, requiring only optimization of the linker length connecting the two domains. This design strategy provides a complementary approach to that of designing new protein–protein interfaces, which produce rigid protein cages (16, 17). Also, because conformational dynamics are important for the biological function of many proteins, by maintaining a looser association between subunits, the potential for interfering with the protein’s biological activity is minimized. We consider that the simplicity and generality of this approach may confer advantages for many applications in synthetic biology, such as construction of enzyme nanoreactors, encapsulation of protein cargos, targeted drug delivery, and polyvalent display of epitopes, where atomic-level precision is not necessary.

The design strategy is inherently modular, and one can imagine that, by combining proteins and coiled-coil domains with different symmetries, a variety of cages with different geometries could be constructed. Coiled-coil designs have been described in which oligomerization has been coupled to events such as metal binding (46), a redox environment (47), and pH changes (48). Such programmability could be introduced into the design to make cage assembly and disassembly responsive to environmental conditions or specific ligands. In addition, further optimization of the design may be achieved by fine-tuning the coiled-coil interactions to improve the kinetics of assembly to reduce misfolding and the formation of inclusion bodies.

## Materials and Methods

**Construction of Genes Encoding Fusion Proteins.** Codon-optimized genes ligated into the expression vector pET28b were either commercially synthesized or derived from the other constructs using standard techniques. The sequences of the proteins are included in Table S1.

**Protein Expression and Purification.** Expression constructs were transformed into *E. coli* BL21(DE3) cells. Cells were grown in 2xYT medium with 50 mg/L kanamycin at 37 °C. At an OD<sub>600</sub> of 0.8, the temperature was reduced to 18 °C, and, at an OD<sub>600</sub> of 1.0, protein expression was induced by addition of 0.1 mM IPTG; cells were grown for a further 18 h and harvested by centrifugation.

All purification steps were performed on ice or at 4 °C. Cell pellets were resuspended in 50 mM Hepes buffer, pH 7.5, containing 1 M urea, 300 mM NaCl, 50 mM imidazole, 5% (vol/vol) glycerol, SigmaFAST protease inhibitor, and 1 mg/mL lysozyme, and then lysed by sonication. The lysate was clarified by centrifugation at  $48,000 \times g$  for 30 min and injected onto a HisTrap nickel–nitrilotriacetic acid (Ni-NTA) column, washed with several volumes of the same buffer, and eluted with 50 mM Hepes buffer, pH 7.5, containing 300 mM NaCl, 500 mM imidazole, and 5% (vol/vol) glycerol. Fractions containing proteins of interest were pooled, dialyzed against 25 mM Hepes buffer, pH 7.5, containing 100 mM NaCl and 2 mM EDTA, concentrated by ultrafiltration, and further purified by SEC on a Superose 6 300/10 column equilibrated in the same buffer. Fractions containing proteins of the desired oligomerization state were pooled and further concentrated for analysis.

**AUC.** Sedimentation velocity analysis was performed using a Beckman Proteome Lab XL-I analytical ultracentrifuge (Beckman Coulter) equipped with an AN60TI rotor. Samples were dialyzed against 25 mM Hepes buffer, pH 7.5, containing 100 mM sodium chloride and 1 mM EDTA. The hydrodynamic behavior of the various proteins was analyzed at a protein concentration with initial absorptions of 0.2 at 280 nm. Samples were loaded into precooled standard sector-shaped, two-channel Epon centerpieces with 1.2-cm path length, and allowed to equilibrate at 6 °C for 2 h in the nonspinning rotor before sedimentation. Proteins were sedimented at  $94,350 \times g$ . Absorbance data were collected at a wavelength of 280 nm. Sedimentation velocity data were analyzed by 2DSA using the finite element modeling module provided with the Ultrascan III software ([www.ultrascan.uthscsa.edu](http://www.ultrascan.uthscsa.edu)). Confidence levels for statistics were derived from 2DSA data refinement using a genetic algorithm followed by 50 Monte Carlo simulations. Calculations were performed on the UltraScan LIMS cluster at the Bioinformatics Core Facility, University of Texas Health Science Center at San Antonio.

**Native MS.** After SEC, samples were concentrated to ~5 mg/mL and then buffer-exchanged into 200 mM ammonium acetate, pH 7.0, using a Bio-spin P30 column (Bio-Rad, Inc.); 2–3  $\mu\text{L}$  of the sample was loaded into glass capillary (approximate o.d. of 1.5–1.8 mm and wall thickness of 0.2 mm) before mounting to the source of an Exactive Plus EMR mass spectrometer (Thermo Fisher Scientific). An electrospray voltage of 1.2 kV was applied to the sample using a platinum wire inserted into the capillary, the source temperature was set to 175 °C, in-source CID was minimized to 1 V or 2 V, HCD was 20 V, the resolution was set to 17,500, and other instrument parameters were set as described previously (43).

**EM Imaging.** Protein complex samples were first screened by negative stain EM. The concentrated samples were diluted to ~0.02 mg/mL and fixed on a grid using conventional negative staining procedures (49). Imaging was performed at room temperature with a Morgagni 268(D) transmission electron microscope (FEI Co.) equipped with a tungsten filament operated at an acceleration voltage of 100 kV and a mounted Orius SC200W CCD camera (Gatan).

For cryo-EM, 3  $\mu\text{L}$  of concentrated sample solution was adsorbed on a glow-discharged Quantifoil grid (R2/2 200 mesh) and vitrified using a Vitrobot (FEI Mark IV). The sample was imaged on a Tecnai TF20 transmission electron microscope (FEI Co.) equipped with a field emission electron gun operated at 200 kV. Images were recorded at a magnification of  $41,667\times$  on a Gatan K2 Summit camera, and binned ( $2 \times 2$  pixels), resulting in a pixel size of 4.4 Å on the specimen level. All of the images were acquired using a low-dose procedure to minimize radiation damage to the samples, with a defocus value of 2–4  $\mu\text{m}$ .

**The 2D Classifications.** A total of 44,856 particle images representing protein cages were manually excised using RELION (50). The contrast transfer function parameters were determined and corrected through e2workflow.py (51). Particles were then subjected to reference-free alignment, classification, and averaging using ISAC. The full set of candidate class averages is shown in Fig. S3. Fully

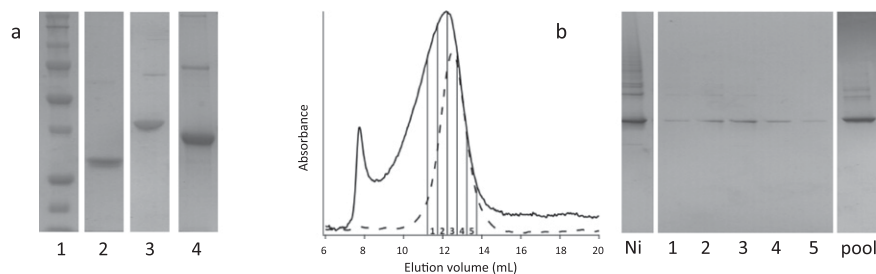
assembled and well-defined class average images were selected to generate the initial mode using program `e2initialmodel.py` (Fig. S4A). Then, 34,980 particles were extracted from those selected classes for 3D reconstruction using RELION. Initial mode was filtered to 60-Å resolution, and then subjected to 3D auto refinement with initial angular sampling at 7.5°. Octahedral (O) symmetry was enforced during reconstruction, and the final map of the protein cage was produced with an indicated resolution of 17 Å at the 0.5 level of Fourier shell correlation (Fig. S4B). The crystal structure of the esterase (PDB 1Z0I) was first manually docked in the map with the C terminus in close proximity to the fourfold axis. The fitting was then refined using the “fit in map” routine in CHIMERA (52). Map visualization, rendering, and figure generation were performed using CHIMERA.

- Kuhn RJ, Rossmann MG (2005) Structure and assembly of icosahedral enveloped RNA viruses. *Adv Virus Res* 64:263–284.
- Zhou ZH, McCarthy DB, O'Connor CM, Reed LJ, Stoops JK (2001) The remarkable structural and functional organization of the eukaryotic pyruvate dehydrogenase complexes. *Proc Natl Acad Sci USA* 98(26):14802–14807.
- Tanaka S, Sawaya MR, Yeates TO (2010) Structure and mechanisms of a protein-based organelle in *Escherichia coli*. *Science* 327(5961):81–84.
- Jutz G, van Rijn P, Santos Miranda B, Böker A (2015) Ferritin: A versatile building block for bionanotechnology. *Chem Rev* 115(4):1653–1701.
- Reisler E, Egelman EH (2007) Actin structure and function: What we still do not understand. *J Biol Chem* 282(50):36133–36137.
- Janke C (2014) The tubulin code: Molecular components, readout mechanisms, and functions. *J Cell Biol* 206(4):461–472.
- Krishna KA, Rao GV, Rao KR (2007) Chaperonin GroEL: Structure and reaction cycle. *Curr Protein Pept Sci* 8(5):418–425.
- Papapostolou D, Howorka S (2009) Engineering and exploiting protein assemblies in synthetic biology. *Mol Biosyst* 5(7):723–732.
- King NP, Lai Y-T (2013) Practical approaches to designing novel protein assemblies. *Curr Opin Struct Biol* 23(4):632–638.
- Lai Y-T, King NP, Yeates TO (2012) Principles for designing ordered protein assemblies. *Trends Cell Biol* 22(12):653–661.
- Channon K, Bromley EHC, Woolfson DN (2008) Synthetic biology through biomolecular design and engineering. *Curr Opin Struct Biol* 18(4):491–498.
- Uchida M, Qazi S, Edwards E, Douglas T (2015) Use of protein cages as a template for confined synthesis of inorganic and organic nanoparticles. *Methods Mol Biol* 1252:17–25.
- Patterson DP, Rynda-Apple A, Harmsen AL, Harmsen AG, Douglas T (2013) Biomimetic antigenic nanoparticles elicit controlled protective immune response to influenza. *ACS Nano* 7(4):3036–3044.
- Padilla JE, Colovos C, Yeates TO (2001) Nanohedra: Using symmetry to design self-assembling protein cages, layers, crystals, and filaments. *Proc Natl Acad Sci USA* 98(5):2217–2221.
- Yeates TO, Padilla JE (2002) Designing supramolecular protein assemblies. *Curr Opin Struct Biol* 12(4):464–470.
- King NP, et al. (2014) Accurate design of co-assembling multi-component protein nanomaterials. *Nature* 510(7503):103–108.
- King NP, et al. (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336(6085):1171–1174.
- Fletcher JM, et al. (2013) Self-assembling cages from coiled-coil peptide modules. *Science* 340(6132):595–599.
- Brodin JD, et al. (2012) Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. *Nat Chem* 4(5):375–382.
- Lanci CJ, et al. (2012) Computational design of a protein crystal. *Proc Natl Acad Sci USA* 109(19):7304–7309.
- Carlson JCT, et al. (2006) Chemically controlled self-assembly of protein nanorings. *J Am Chem Soc* 128(23):7630–7638.
- Ringler P, Schulz GE (2003) Self-assembly of proteins into designed networks. *Science* 302(5642):106–109.
- Lai Y-T, Tsai K-L, Sawaya MR, Asturias FJ, Yeates TO (2013) Structure and flexibility of nanoscale protein cages designed by symmetric self-assembly. *J Am Chem Soc* 135(20):7738–7743.
- Kobayashi N, et al. (2015) Self-assembling nano-architectures created from a protein nano-building block using an intermolecularly folded dimeric de novo protein. *J Am Chem Soc* 137(35):11285–11293.
- Huard DJE, Kane KM, Tezcan FA (2013) Re-engineering protein interfaces yields copper-inducible ferritin cage assembly. *Nat Chem Biol* 9(3):169–176.
- Lai Y-T, Cascio D, Yeates TO (2012) Structure of a 16-nm cage designed by using protein oligomers. *Science* 336(6085):1129.
- Lai Y-T, et al. (2014) Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* 6(12):1065–1071.
- Raman S, Machaidze G, Lustig A, Aebi U, Burkhard P (2006) Structure-based design of peptides that self-assemble into regular polyhedral nanoparticles. *Nanomedicine (Lond)* 2(2):95–102.
- Usui K, et al. (2009) Nanoscale elongating control of the self-assembled protein filament with the cysteine-introduced building blocks. *Protein Sci* 18(5):960–969.
- Patterson DP, et al. (2014) Characterization of a highly flexible self-assembling protein system designed to form nanocages. *Protein Sci* 23(2):190–199.
- Patterson DP, Desai AM, Holl MMB, Marsh ENG (2011) Evaluation of a symmetry-based strategy for assembling protein complexes. *RSC Advances* 1(6):1004–1012.
- Elmi F, et al. (2005) Stereoselective esterase from *Pseudomonas putida* IFO12996 reveals alpha/beta hydrolase folds for D-beta-acetylthioisobutyric acid synthesis. *J Bacteriol* 187(24):8470–8476.
- Lupas AN, Gruber M (2005) The structure of alpha-helical coiled coils. *Adv Protein Chem* 70:37–78.
- Fletcher JM, et al. (2012) A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology. *ACS Synth Biol* 1(6):240–250.
- Thomas F, Boyle AL, Burton AJ, Woolfson DN (2013) A set of de novo designed parallel heterodimeric coiled coils with quantified dissociation constants in the micromolar to sub-nanomolar regime. *J Am Chem Soc* 135(13):5161–5166.
- Negron C, Keating AE (2014) A set of computationally designed orthogonal antiparallel homodimers that expands the synthetic coiled-coil toolkit. *J Am Chem Soc* 136(47):16544–16556.
- Zaccari NR, et al. (2011) A de novo peptide hexamer with a mutable channel. *Nat Chem Biol* 7(12):935–941.
- Das R, Baker D (2008) Macromolecular modeling with Rosetta. *Annu Rev Biochem* 77:363–382.
- Demeler B, Saber H, Hansen JC (1997) Identification and interpretation of complexity in sedimentation velocity boundaries. *Biophys J* 72(1):397–407.
- Demeler B, van Holde KE (2004) Sedimentation velocity analysis of highly heterogeneous systems. *Anal Biochem* 335(2):279–288.
- Demeler B (2005) UltraScan - A comprehensive data analysis software package for analytical ultracentrifugation experiments. *Analytical Ultracentrifugation: Techniques and Methods*, eds Scott DJ, Harding SE, Rowe AJ (R Soc Chem, London), pp 210–230.
- Bosma HJ, De Kok A, Van Markwijk BW, Veege C (1984) The size of the pyruvate dehydrogenase complex of *Azotobacter vinelandii*. Association phenomena. *Eur J Biochem* 140(2):273–280.
- McKay AR, Ruotolo BT, Ilag LL, Robinson CV (2006) Mass measurements of increased accuracy resolve heterogeneous populations of intact ribosomes. *J Am Chem Soc* 128(35):11433–11442.
- Rose RJ, Damoc E, Denisov E, Makarov A, Heck AJR (2012) High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. *Nat Methods* 9(11):1084–1086.
- Yang Z, Fang J, Chittuluru J, Asturias FJ, Penczek PA (2012) Iterative stable alignment and clustering of 2D transmission electron microscope images. *Structure* 20(2):237–247.
- Marsh ENG, DeGrado WF (2002) Noncovalent self-assembly of a heterotetrameric diiron protein. *Proc Natl Acad Sci USA* 99(8):5150–5154.
- Zhou NE, Kay CM, Hodges RS (1993) Disulfide bond contribution to protein stability: Positional effects of substitution in the hydrophobic core of the two-stranded alpha-helical coiled-coil. *Biochemistry* 32(12):3178–3187.
- Zimenkov Y, et al. (2006) Rational design of a reversible pH-responsive switch for peptide self-assembly. *J Am Chem Soc* 128(21):6770–6771.
- Ohi M, Li Y, Cheng Y, Walz T (2004) Negative staining and image classification: Powerful tools in modern electron microscopy. *Biol Proced Online* 6:23–34.
- Scheres SHW (2012) RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* 180(3):519–530.
- Tang G, et al. (2007) EMAN2: An extensible image processing suite for electron microscopy. *J Struct Biol* 157(1):38–46.
- Pettersen EF, et al. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612.

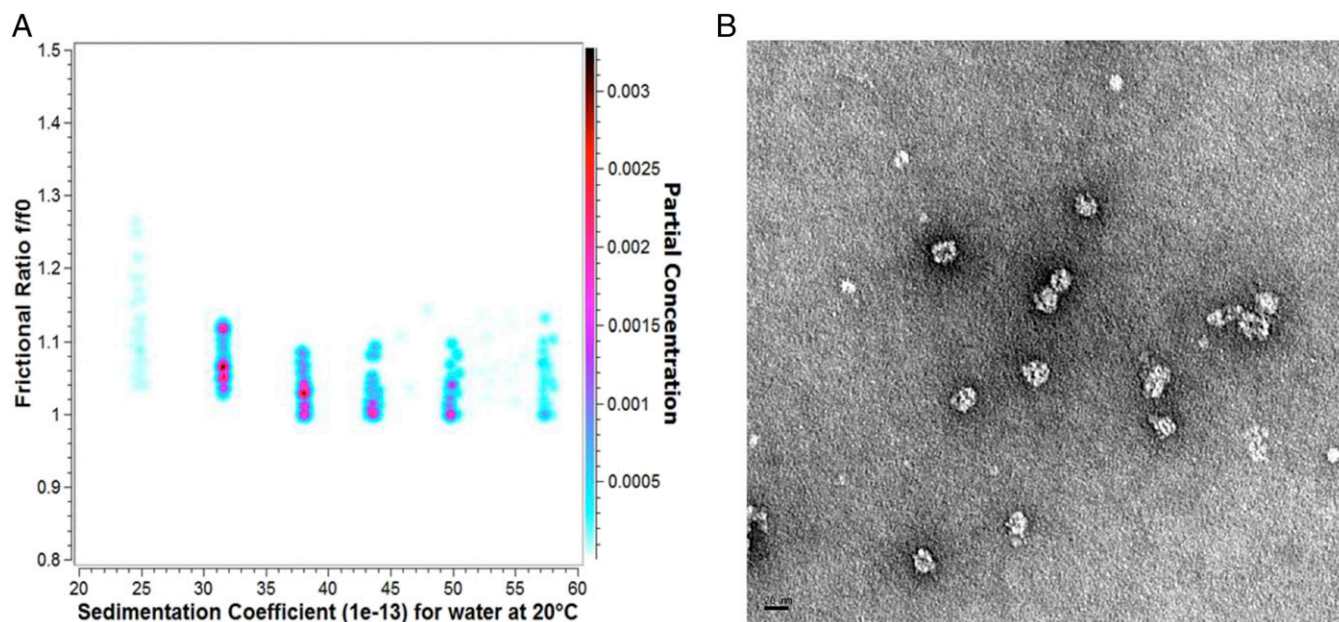


# Supporting Information

Sciore et al. 10.1073/pnas.1606013113



**Fig. S1.** (A) SDS PAGE of proteins. Lane 1, protein standards; lane 2, unmodified esterase; lane 3, Oct-4; and lane 4, Oct-2. (B, Left) SEC of Oct-4 after purification on Ni-NTA resin (solid trace). Fractions 1–5 were analyzed by native PAGE, pooled, and rechromatographed (dashed trace). (Right) Analysis of SEC fractions by native PAGE. Lanes on the gel are: Ni, Oct-4 after purification on Ni-NTA resin; lanes 1–5, fractions 1–5; and pool, pooled material after SEC.



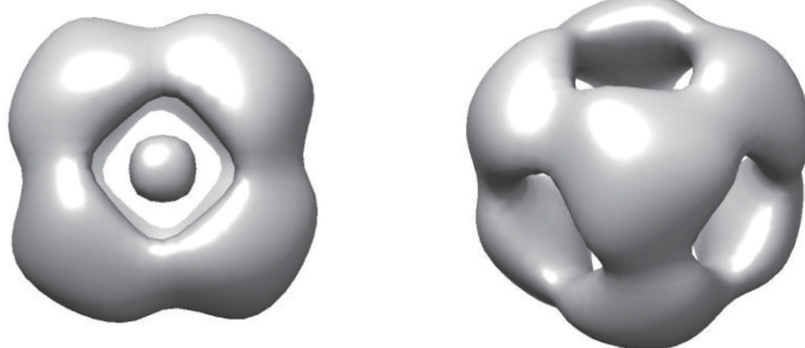
**Fig. S2.** Further characterization of Oct-2. (A) A 2DSA of Oct-2. The protein forms multiple species characterized by sedimentation coefficients that are larger than expected for an octahedral cage. The low frictional ratios are consistent with the formation of globular complexes. (B) Negative stain EM of Oct-2. The images indicate that the protein assembles into a range of particle sizes, but no symmetry is apparent in the images, in contrast to the particles formed by Oct-4 (Fig. 3D). (Scale bar, 20 nm.)



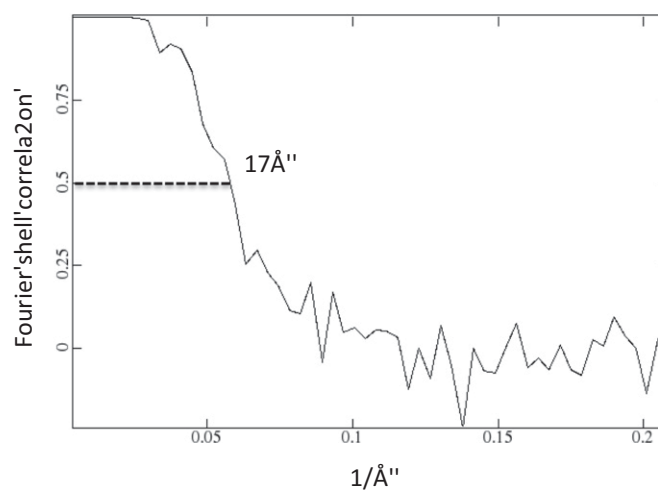


3'fold'

a



**b**



**Fig. S4.** (A) Initial electron density model used in 3D reconstruction of Oct-4 from cryo-EM data. Model is shown viewed along threefold and fourfold symmetry axes. (B) Estimation of resolution of the reconstructed model of Oct-4. The final map of the protein cage was produced with an indicated resolution of 17 Å at the 0.5 level of Fourier shell correlation.

**Table S1. Amino acid sequences of proteins used in this study**

Protein	Amino acid sequence
Trimeric esterase	MGSSHHHHHSSGLVPRGSHMSYVTTKDGVIIFYKDWGPRDAPVIHFHHGWPLSADDWDAQLLFFLAHGYRVVAHRRRGHGRSSQVWDGH-DMDHYADDVAAVVAHLGIQGAHVHGHSTGGGEVVRIMARHPEDKVAKAVLIAAVPPLMVQTPGNPGGLPKSVFDGFGAQVASNRAQFY-RDVPAGPFYGYNRPGEASEGIIGNWWRQGMIGSAKAHYDGIVAFSQTDFTEDLKGIQQPVLVMHGDDQIVPYENSGVLSAKLLPENG-ALKTYKGYPHGMPPTHADVINADLLAFIRS
Oct-4	MGSSHHHHHSSGLVPRGSHMSYVTTKDGVIIFYKDWGPRDAPVIHFHHGWPLSADDWDAQLLFFLAHGYRVVAHRRRGHGRSSQVWDGH-DMDHYADDVAAVVAHLGIQGAHVHGHSTGGGEVVRIMARHPEDKVAKAVLIAAVPPLMVQTPGNPGGLPKSVFDGFGAQVASNRAQFY-RDVPAGPFYGYNRPGEASEGIIGNWWRQGMIGSAKAHYDGIVAFSQTDFTEDLKGIQQPVLVMHGDDQIVPYENSGVLSAKLLPENG-ALKTYKGYPHGMPPTHADVINADLLAFIRSGTGLAAIKQELAAIRSELAAIKHELAAIKQE
Oct-3	MGSSHHHHHSSGLVPRGSHMSYVTTKDGVIIFYKDWGPRDAPVIHFHHGWPLSADDWDAQLLFFLAHGYRVVAHRRRGHGRSSQVWDGH-DMDHYADDVAAVVAHLGIQGAHVHGHSTGGGEVVRIMARHPEDKVAKAVLIAAVPPLMVQTPGNPGGLPKSVFDGFGAQVASNRAQFY-RDVPAGPFYGYNRPGEASEGIIGNWWRQGMIGSAKAHYDGIVAFSQTDFTEDLKGIQQPVLVMHGDDQIVPYENSGVLSAKLLPENG-ALKTYKGYPHGMPPTHADVINADLLAFIRSGTGLAAIKQELAAIRSELAAIKHELAAIKQE
Oct-2	MGSSHHHHHSSGLVPRGSHMSYVTTKDGVIIFYKDWGPRDAPVIHFHHGWPLSADDWDAQLLFFLAHGYRVVAHRRRGHGRSSQVWDGH-DMDHYADDVAAVVAHLGIQGAHVHGHSTGGGEVVRIMARHPEDKVAKAVLIAAVPPLMVQTPGNPGGLPKSVFDGFGAQVASNRAQFY-RDVPAGPFYGYNRPGEASEGIIGNWWRQGMIGSAKAHYDGIVAFSQTDFTEDLKGIQQPVLVMHGDDQIVPYENSGVLSAKLLPENG-ALKTYKGYPHGMPPTHADVINADLLAFIRSGTLAAIKQELAAIRSELAAIKHELAAIKQE

The flexible linker region is shown in blue, and the coiled-coil sequence is in red.

**Table S2. Hydrodynamic parameters for protein assemblies formed by Oct-4 determined by sedimentation velocity AUC**

Species	Sedimentation coefficient, S	Molecular weight, kDa	Frictional Ratio ( $f/f_0$ )	Partial concentration, %
Solute 1	17.6 ± 0.1	886 ± 14	1.89 ± 0.02	73.3
Solute 2	22.1 ± 0.07	489 ± 26	1.01 ± 0.04	18.5
Solute 3	27.7 ± 0.3	728 ± 114	1.05 ± 0.1	4.5
Solute 4	37.3 ± 0.2	1,145 ± 194	1.06 ± 0.09	2.3

For details, see *AUC of Oct-4*.

**Table S3. Hydrodynamic parameters for protein assemblies formed by Oct-2 determined by sedimentation velocity AUC**

Species	Sedimentation coefficient, S	Molecular weight, kDa	Frictional ratio ( $f/f_0$ )	Partial concentration, %
Solute 1	24.8 ± 0.4	649 ± 96	1.09 ± 0.11	2.2
Solute 2	31.5 ± 0.1	905 ± 67	1.07 ± 0.05	21.5
Solute 3	38.0 ± 0.2	1,128 ± 63	1.03 ± 0.04	27.4
Solute 4	43.5 ± 0.3	1,357 ± 93	1.01 ± 0.05	20.3
Solute 5	49.8 ± 0.8	1,681 ± 174	1.02 ± 0.06	13.9
Solute 6	57.3 ± 0.7	2,113 ± 199	1.03 ± 0.07	7.4

For details, see *AUC of Oct-4*.

### Other Supporting Information Files

[Dataset S1 \(TXT\)](#)  
[Dataset S2 \(TXT\)](#)